Estimating and Visualizing Population Genetic Structure for Landscape Genomics

What is Landscape Genomics?

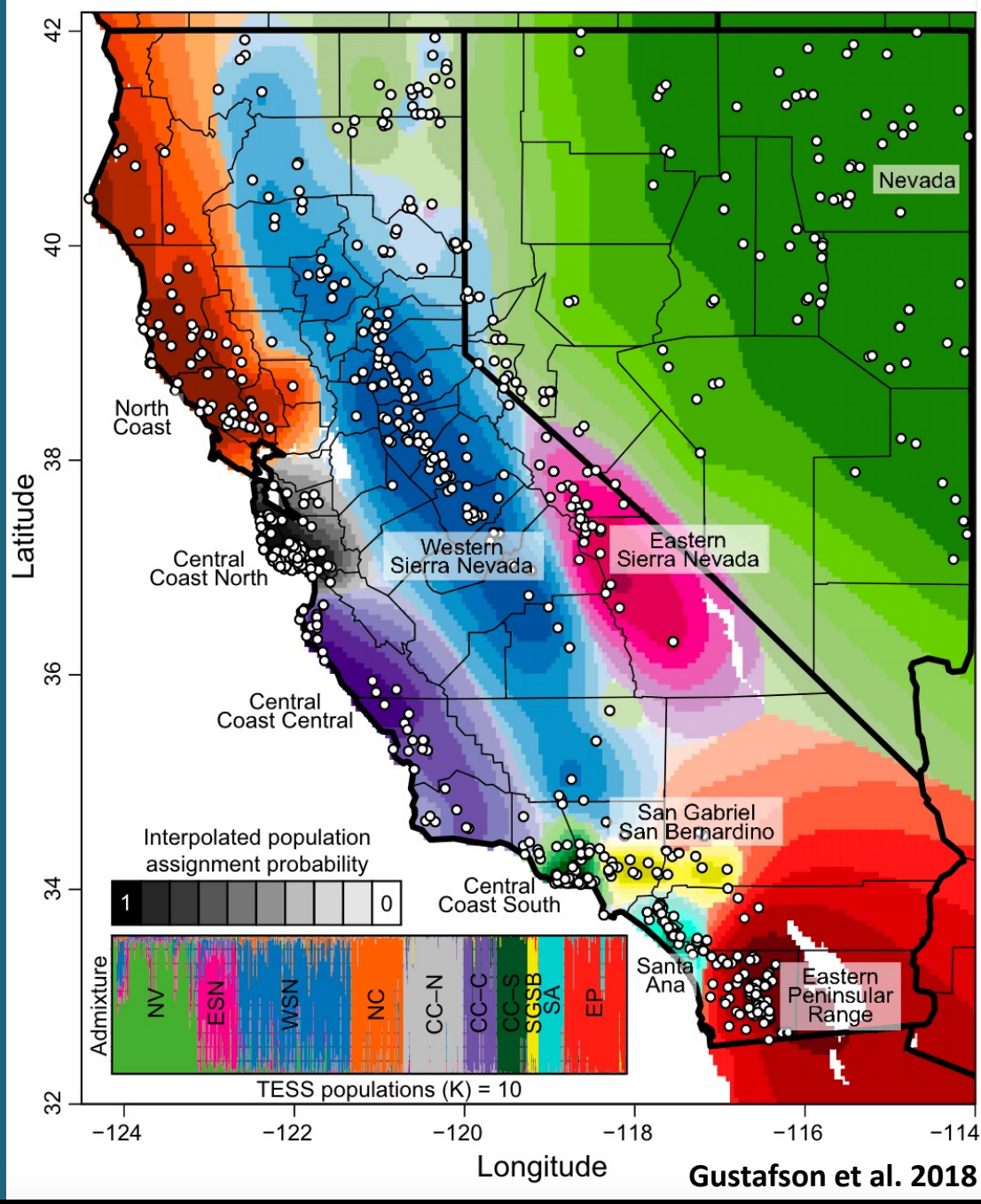
Study of the the interaction between landscapes and genetic variation

Manel et al. 2003

- Introduced the concept of landscape genetics, bridging landscape ecology and population genetics
- Reviewed and described research methodologies for studying the interaction between landscapes and genetic variation

Manel et al. 2013

- Landscape genomics emerging from landscape genetics as technology improved



Gustafson et al. 2018

Why Include Spatial Data?

Genetic data isn't isolated

- It's intricately woven into the fabric of geography

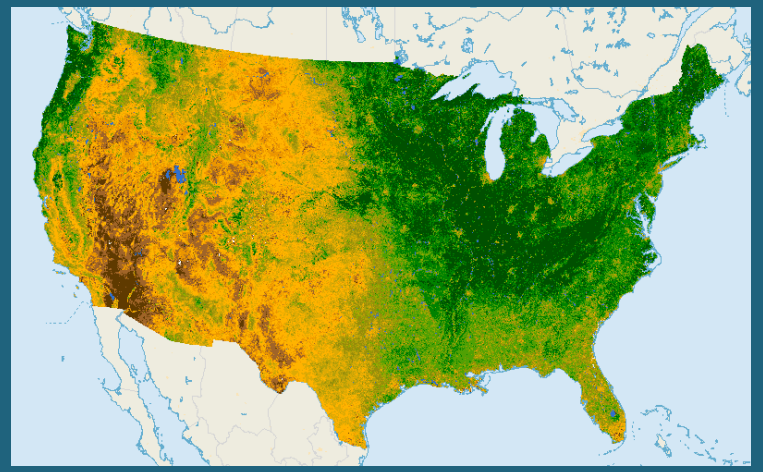
Populations are extensions of their environment

- The environmental characteristics of a landscape are strong drivers of genetic variation and evolutionary processes

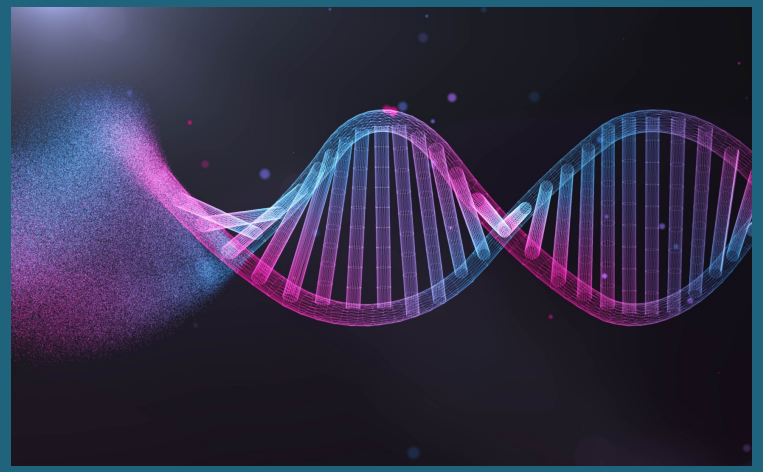
To better understand a species' evolutionary history and trajectory, we must consider the spatial dimensions of genetic information

Types of Data

Geographic Data



Genomic Data



Environmental Data

Geographic Data

Geographic Information Systems (GIS)

- Framework for gathering, managing, and analyzing spatial or geographic data
- Analyzes spatial relationships to uncover patterns and trends
- Can be leveraged to analyze genomic data

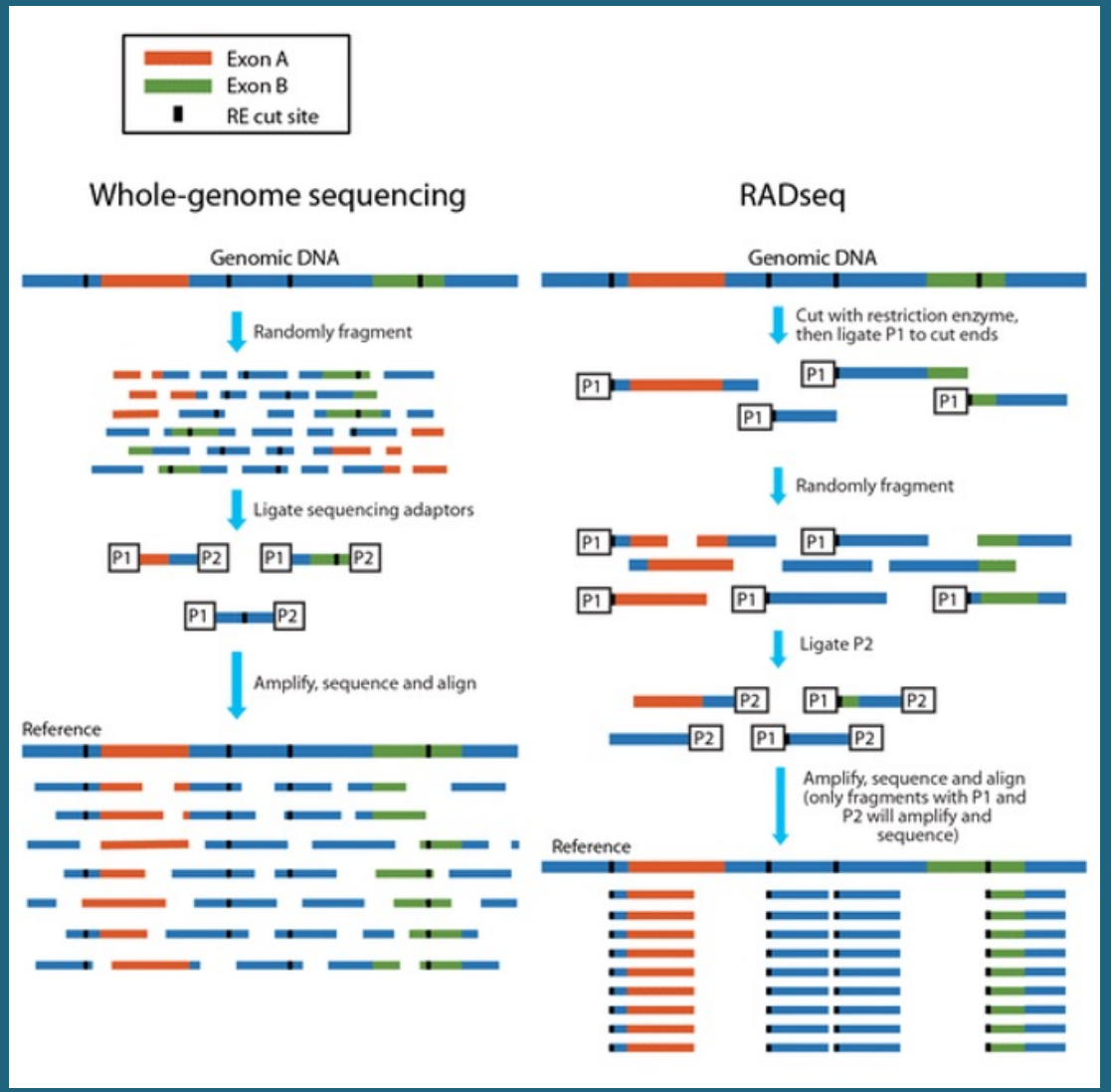
Examples of useful geospatial data:

- Telemetry/tracking Data
- Camera Traps

Genomic Data

Examples:

- Whole Genome Sequencing (WGS)
- Restriction site-associated DNA sequencing (RADseq)
- Microsatellite Markers



Perry et al. 2017

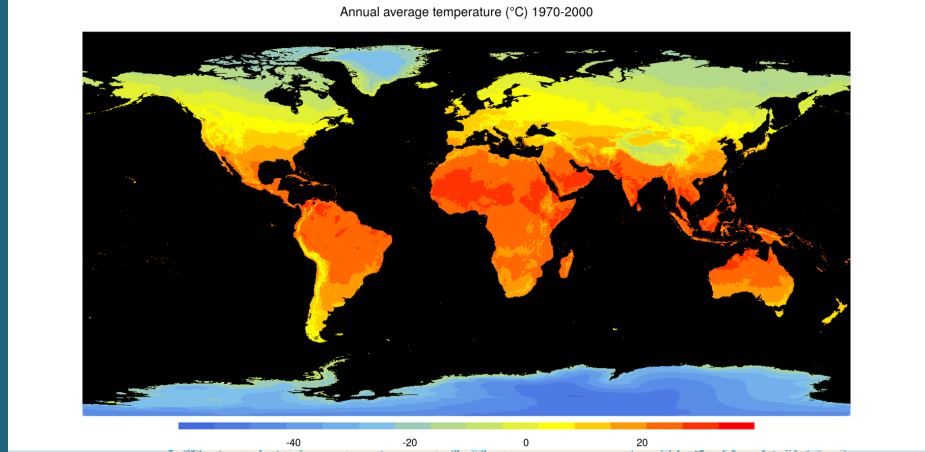
Environmental Data

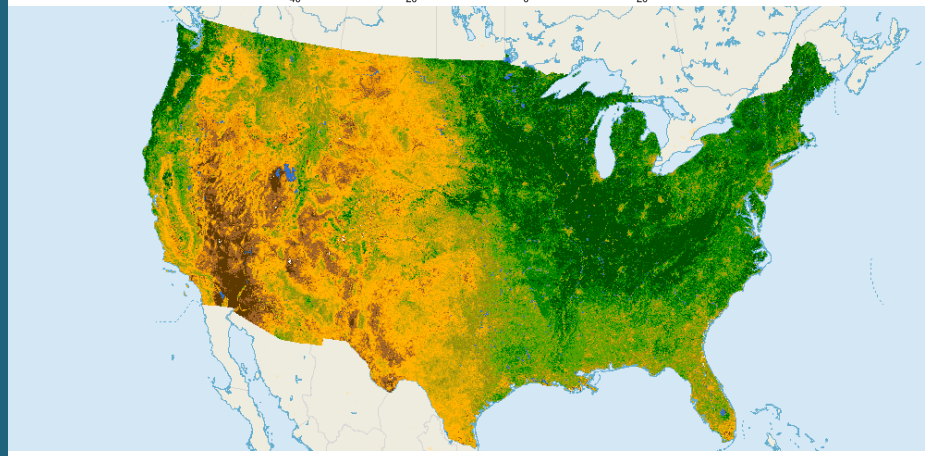
Global Environmental Datasets:

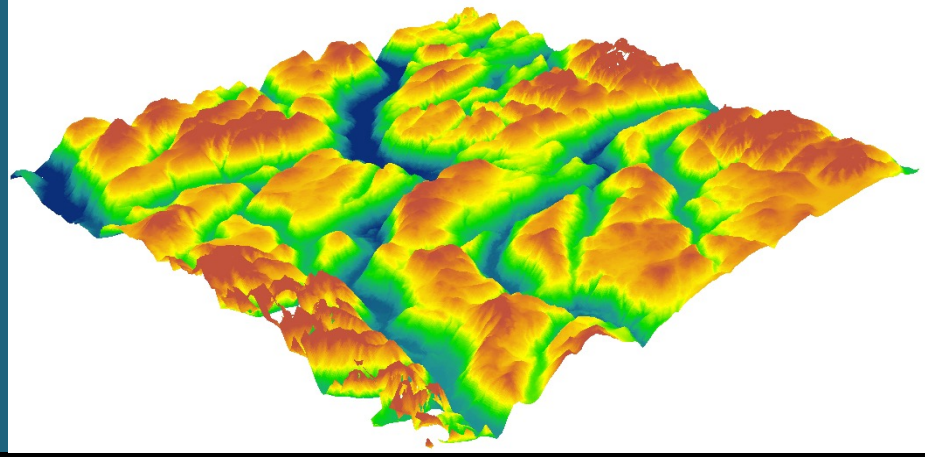
- Source: Government agencies or research institutions
- Can include climate, land use, and ecological variables
- Examples:
  - WorldClim database
  - NASA Landsat data

Regional GIS Databases

- Leverage geographic information systems (GIS) databases that compile various environmental layers, including soil types, topography, and hydrology
- Availability often depends on region







Data Matrix

Geographic Data | **Genomic Data** | **Environmental Data**

Animal_ID	Latitude	Longitude	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	SNP11	SNP12	SNP13	SNP14	SNP15	ELEV	PRECIP	TEMP
1	31.559633	-110.450312	AA	TT	NA	TT	AG	AG	CT	NA	AG	CT	GG	AG	AG	CC	AC	1610	504	15.3303
2	31.8368055	-110.737768	AG	TT	CC	TT	GG	AG	CC	NA	AA	CT	AA	GG	GG	CC	AA	1518	515	16.9441
3	32.44535	-110.495085	AA	TT	CC	AA	GG	AG	CC	NA	AA	TT	AA	AG	GG	NA	AA	914	373	18.3929
4	33.515916	-109.210575	AG	TT	NA	TT	GG	AG	CC	NA	AA	TT	GG	GG	AG	CC	AA	1600	433	11.2909
5	33.539167	-112.558409	GG	TT	NA	AT	AG	GG	CC	NA	AG	CT	AG	AG	GG	CT	AA	887	279	20.7742
6	34.7245611	-111.610024	AG	TT	CT	TT	GG	AA	CT	NA	AG	CT	GG	NA	GG	CT	AA	1825	599	12.9797
7	31.688422	-110.885365	AA	TT	NA	AT	GG	AA	CC	NA	AG	CC	AG	GG	GG	CT	AC	2371	646	13.5509
8	31.79665	-110.78773	AG	TT	NA	TT	GG	AG	CC	NA	GG	CT	AG	GG	GG	CC	AC	1500	562	16.0225
9	31.8939498	-110.769955	AA	TT	CC	TT	NA	AA	CC	NA	AG	CT	AG	AG	AG	CC	CC	1231	491	17.757
10	32.053442	-110.426879	AG	TT	CC	TT	GG	AA	CC	NA	AA	TT	AA	AA	GG	CC	AA	1201	412	17.0669
11	32.132022	-110.519523	AA	TT	NA	AT	GG	AG	TT	NA	GG	TT	GG	AA	GG	CC	AA	2184	554	14.9285
12	32.395097	-110.811021	AA	TT	NA	AA	GG	AA	CT	NA	AA	TT	GG	AG	GG	CC	AA	1623	590	15.247
13	32.412741	-110.730386	AG	TT	NA	AA	GG	AG	CC	NA	AA	TT	AG	AG	GG	CC	AC	2375	647	13.5843
14	32.42187	-111.48162	AA	TT	NA	TT	GG	AG	CC	NA	AG	CT	AG	GG	GG	CC	AC	847	297	20.8507
15	32.5117	-111.0752	GG	TT	NA	AT	GG	AA	CC	NA	AG	TT	AG	AG	GG	CC	AA	1052	474	18.4648
16	32.5509	-111.05443	AA	TT	NA	AT	GG	AG	TT	NA	AA	CC	AG	AG	AG	CC	AA	1167	474	18.4648
17	33.082846	-111.317039	AG	CT	NA	AT	GG	AG	CC	NA	GG	CT	AG	AA	GG	CC	AA	479	284	21.4194
18	33.138323	-112.80343	AA	TT	NA	AT	GG	GG	CC	NA	AG	CT	GG	GG	AA	CC	CC	369	185	22.8287
19	34.021595	-113.431287	AG	TT	NA	AT	GG	AG	CT	NA	AG	CC	AG	GG	AG	CC	AC	1096	297	19.2441

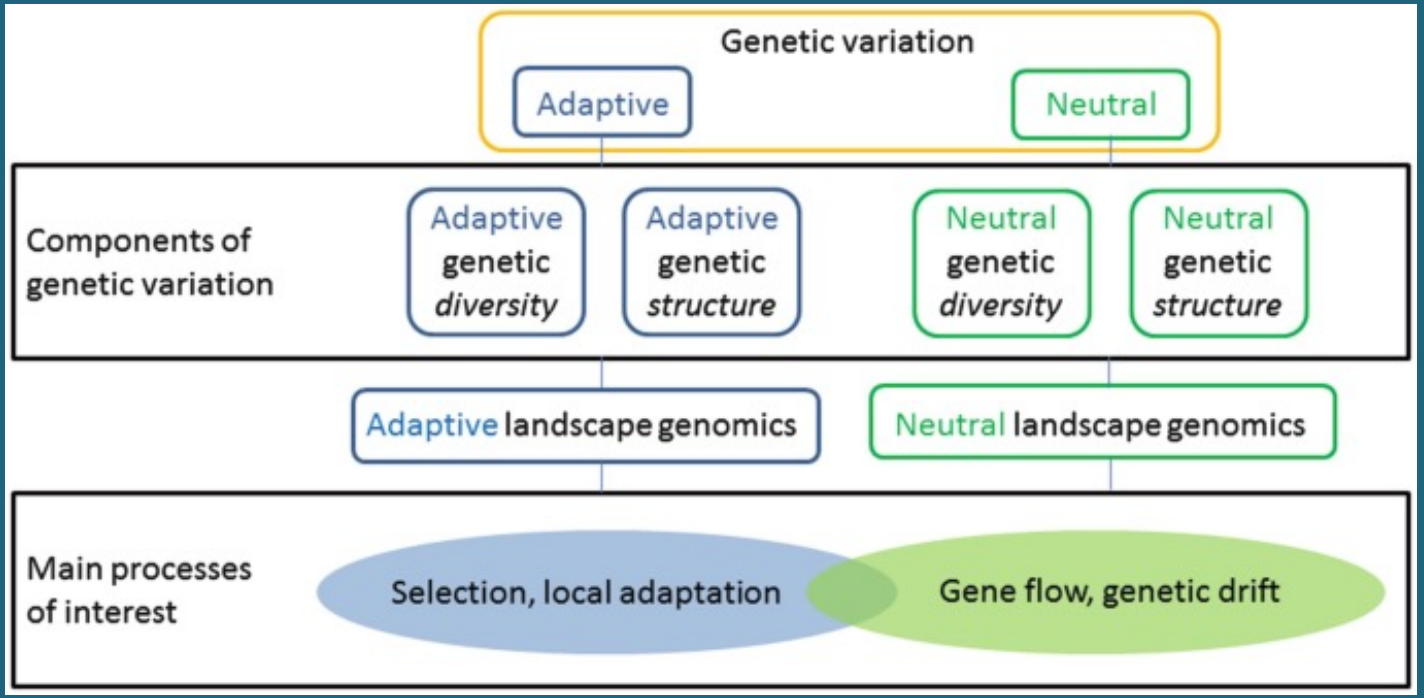
How Can We Use This Data to Understand Evolutionary Processes?

Assess adaptive evolution

- Identifies genetic variations influenced by local environmental factors
- Example: Gene-Environment Associations (GEA)

Assess non-adaptive (neutral) evolution

- Analysis of spatial genetic structure
- Evaluate relationships between individuals to estimate relatedness based on genetic data



Balkenhol et al. 2017

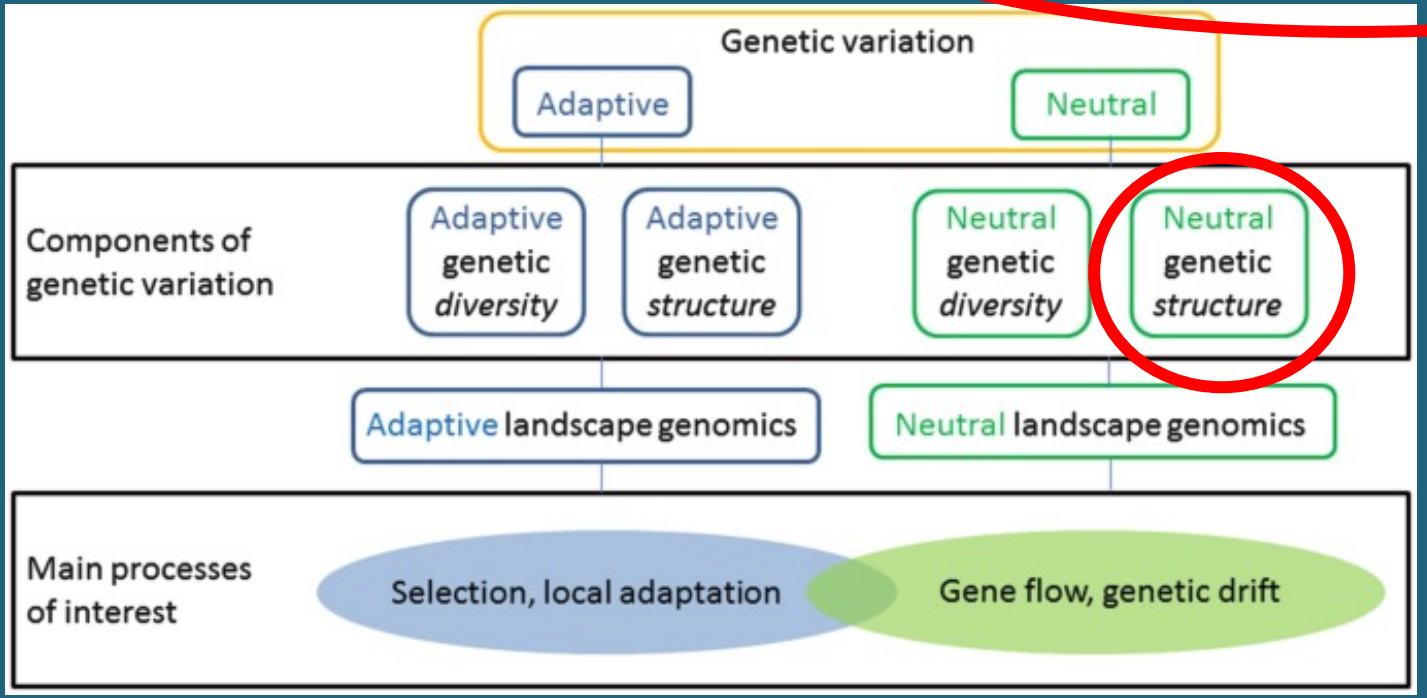
How Can We Use This Data to Understand Evolutionary Processes?

Assess adaptive evolution

- Identifies genetic variations influenced by local environmental factors
- Example: Gene-Environment Associations (GEA)

Assess non-adaptive (neutral) evolution

- Analysis of spatial genetic structure
- Evaluate relationships between individuals to estimate relatedness based on genetic data



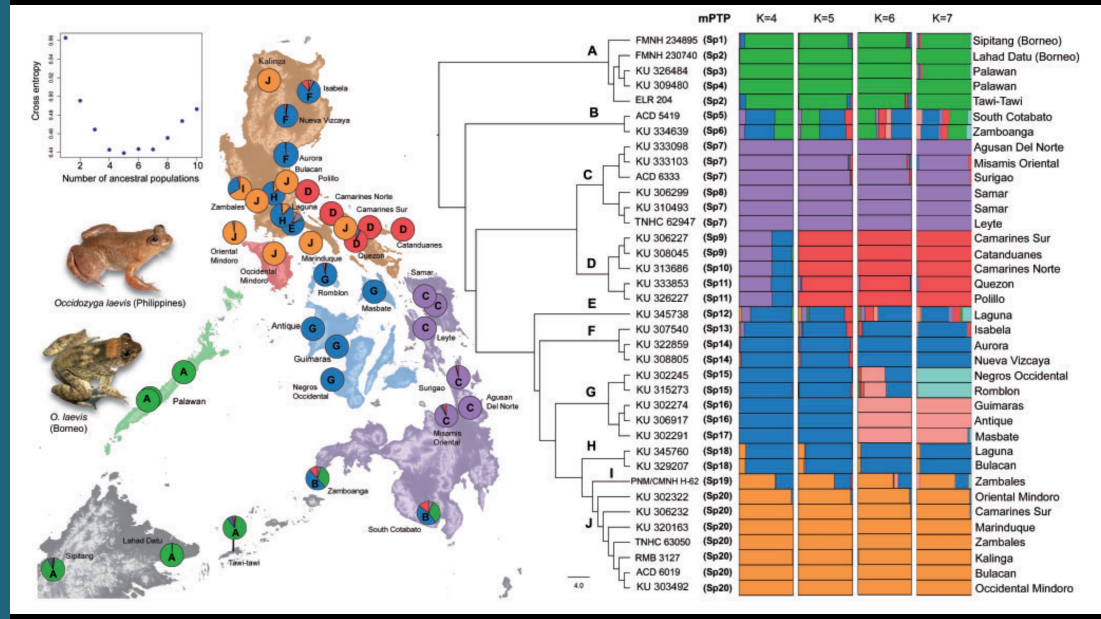
Balkenhol et al. 2017

Analyzing Spatial Genetic Structure

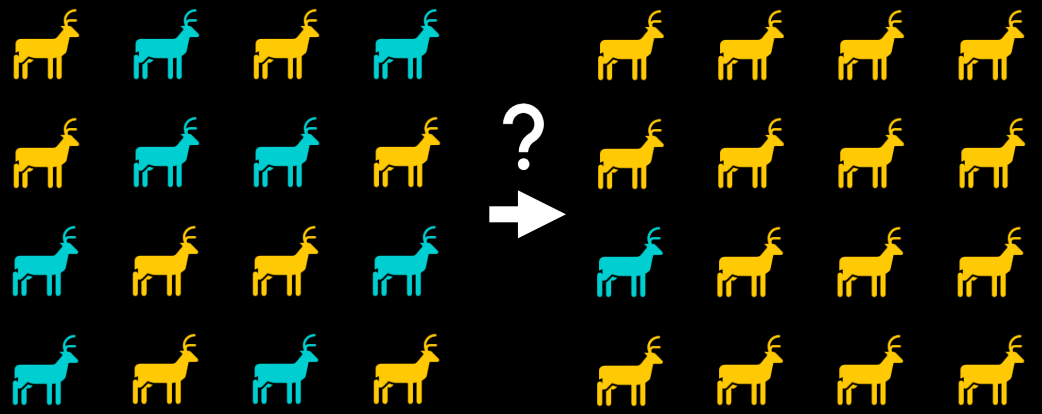
Genetic Structure

- Often measured in terms of **ancestry coefficients**
  - The proportion of individual's genomes that originate from multiple ancestral gene pools

- Genetic structure can mimic selection, despite being driven by non-selective factors
  - Quantifying genetic structure helps incorporate its effects into subsequent analyses focused on adaptations within population



Chan et al. 2022



Analyzing Spatial Genetic Structure

STRUCTURE (Pritchard et al. 2000)

- Uses matrix models and Bayesian statistics to estimate ancestry coefficients
- Stochastic algorithms (like MCMC in STRUCTURE) can be computationally demanding when using large numbers of individuals or loci

Inference of Population Structure Using Multilocus Genotype Data

Jonathan K. Pritchard, Matthew Stephens and Peter Donnelly

Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom

Manuscript received September 23, 1999
Accepted for publication February 18, 2000

TESS3 (Chen et al. 2007)

- Model-free and spatially explicit
- Accounts for geographic distance among individuals to interpolate ancestry coefficients on a landscape

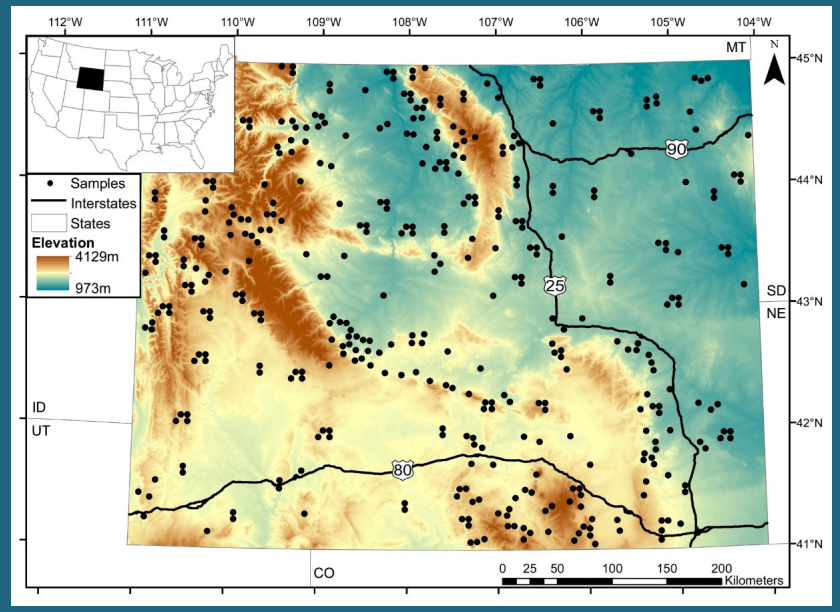
TESS3: fast inference of spatial population structure and genome scans for selection

KEVIN CAYE,* TIMO M. DEIST,* HELENA MARTINS,* OLIVIER MICHEL† and OLIVIER FRANÇOIS*

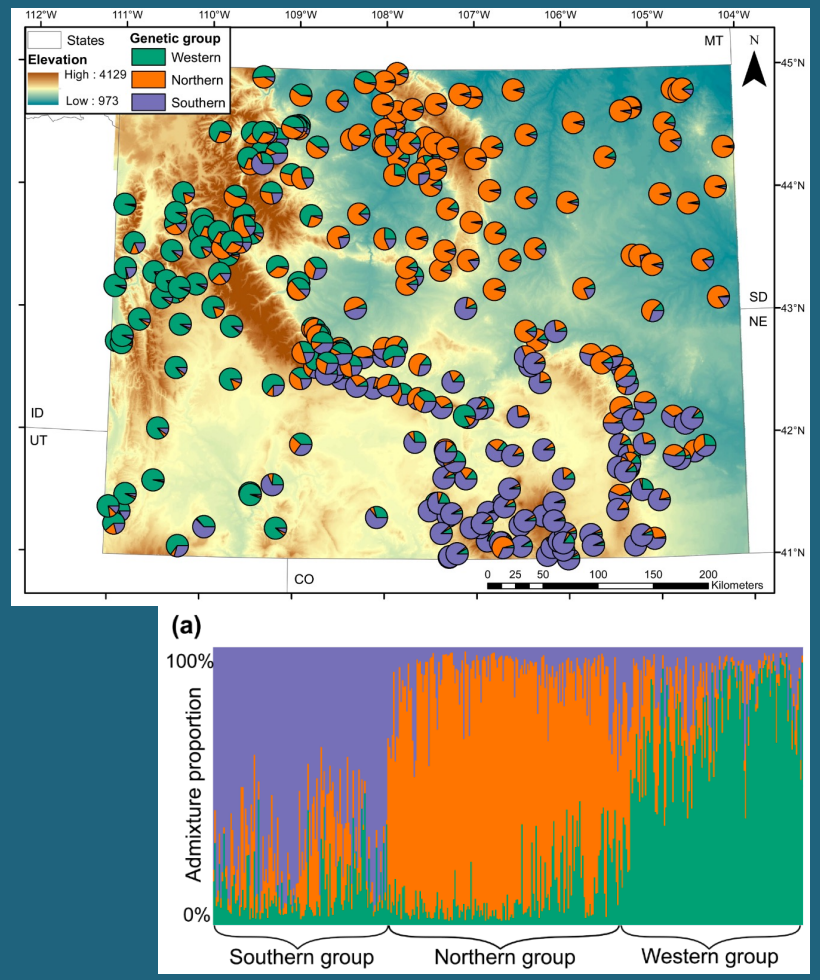
**Centre National de la Recherche Scientifique, Université Grenoble-Alpes, TIMC-IMAG UMR 5525, Grenoble 38042, France,*
†Centre National de la Recherche Scientifique, Université Grenoble-Alpes, GIPSA-lab UMR 5216, Grenoble 38042, France

Analyzing Spatial Genetic Structure

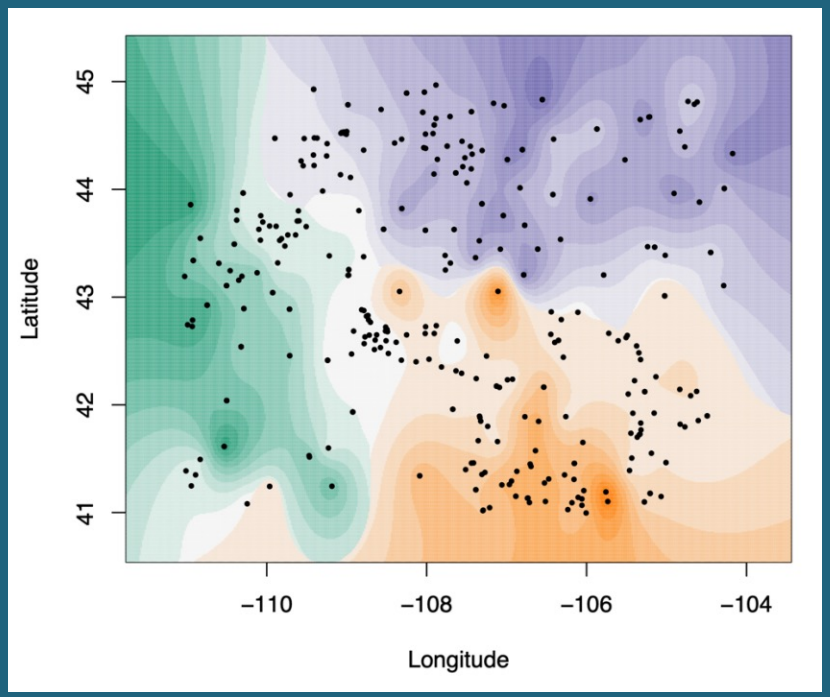
Occurrence Data



STRUCTURE



TESS3



LaCava et al. 2021

TESS3 Demo

TESS3: Estimating Ancestry Coefficients

TESS3 is available as R package `tess3r`

TESS3 estimates ancestry coefficients (Q) using a least-squares minimization algorithm

- Generated when least squares are minimized between X and QG

$$P(X_{il} = j) = \sum_{k=1}^{K} Q_{ik} G_{kl}(j)$$

Probability that individual i carries the genotype j at locus l (e.g. AA)

Based on observed data

Individual ancestry coefficients

Fraction of an individual's genome that originates from ancestral population k

Ancestral genotypic frequencies

Frequency of genotype j at locus l in population k

Caye et al 2016

TESS3: Input Data

We will be using the built-in example data for *Arabidopsis thaliana* (`data.at`)

TESS3 primarily uses the following types of input data:

- Genomic data
  - **Genotype matrix**
- Geographic data
  - **Longitude and latitude**

What is the spatial extent (scale) of this data?

What do rows and columns represent?

How many SNPs are we analyzing?

TESS3: Function and Arguments

Will take 1-15 minutes to run on most laptops

```
tess3.obj <- tess3(X = genotype,
                   coord = coordinates,
                   K = 1:10,
                   method = "projected.ls",
                   ploidy = 1,
                   openMP.core.num = 4)
```

`X`: Genotype matrix
`coord`: Coordinate data

`K`: Number of possible ancestral populations we'd like to compare

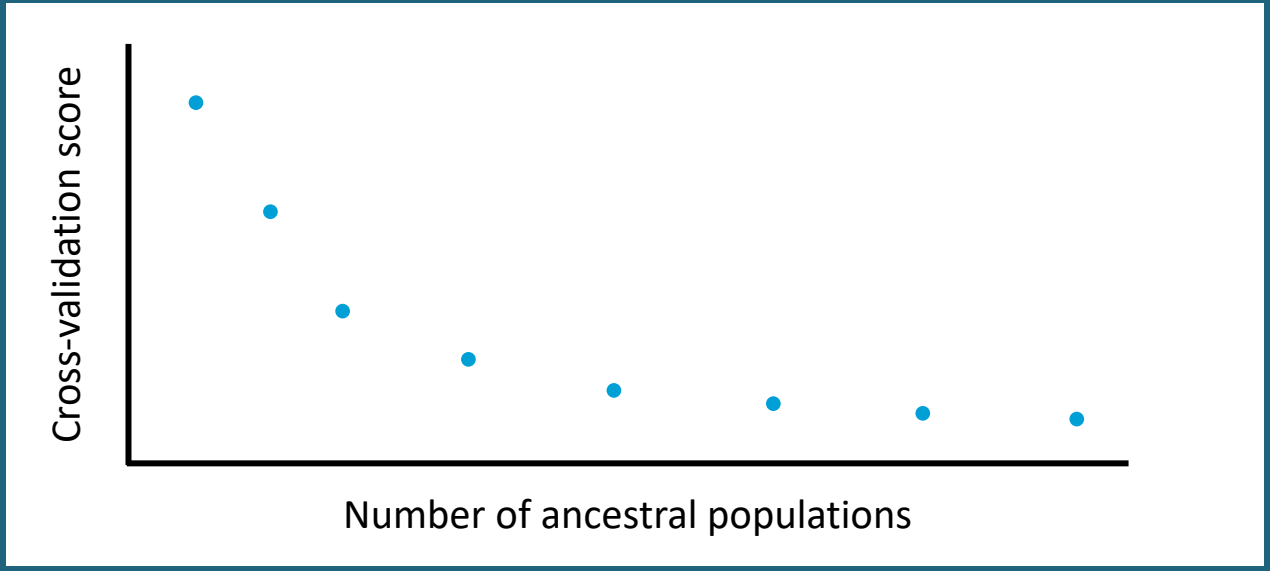
`method`: Least squares algorithm used (default as described is "projected.ls")

`ploidy`: Ploidy of organism

TESS3: Choosing Optimal Number of K

K is determined using cross-entropy cross-validation criterion (Frichot et al. 2014)

- Partitions genotypic matrix into training and test sets
- Compares predicted genotypic frequencies between observed sets at each locus
- Smaller values imply better estimates in TESS3



Somewhat subjective: Optimal K choice observed at plateau in cross-entropy plot

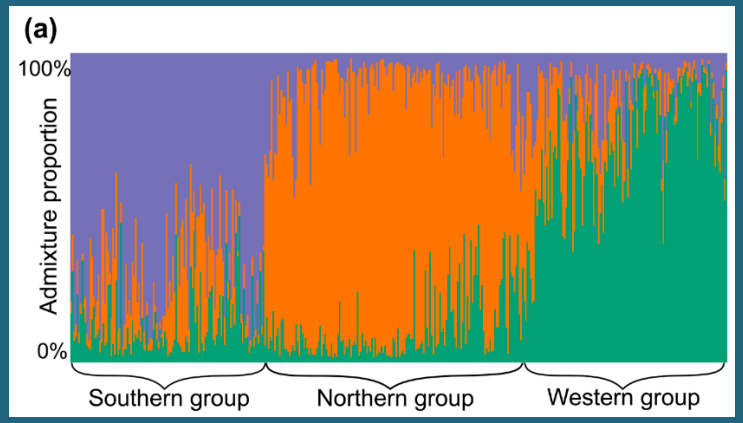
- **Defines major clusters, but not necessarily all hierarchical structuring**

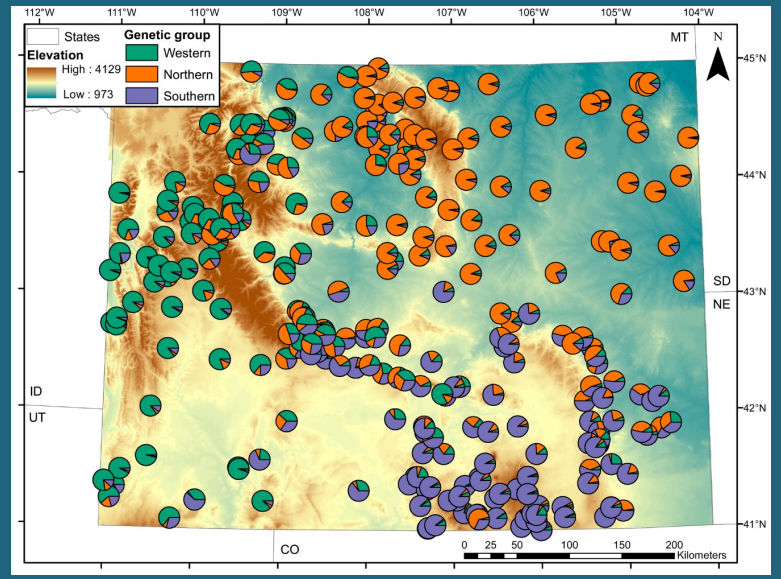
TESS3: Plotting Ancestry Coefficients

STRUCTURE, ADMIXTURE, and similar approaches often represent ancestry coefficients with box plots, where colors represent ancestral populations and columns represent individuals

If geographic data is available, box plots may be converted to pie charts layered on a map

TESS3 output can similarly be represented this way

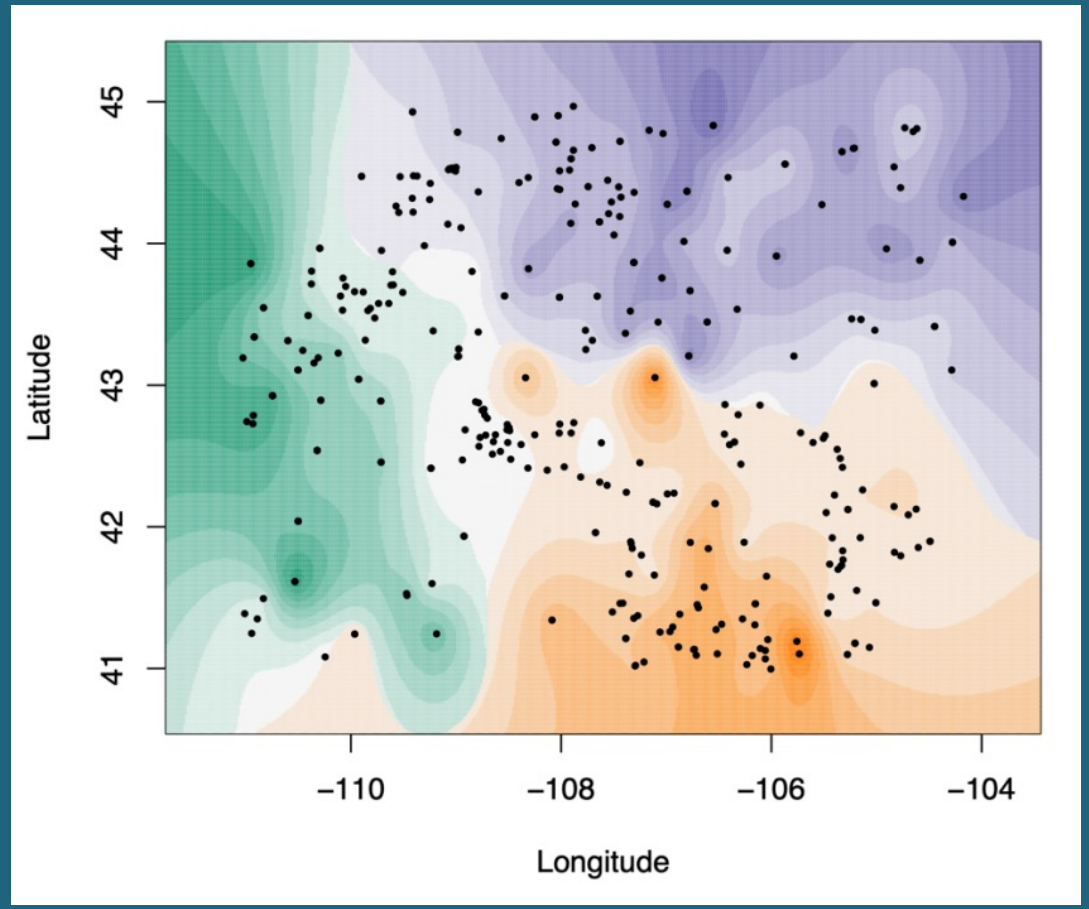




LaCava et al. 2021

TESS3: Plotting Ancestry Coefficients

`plot` can interpolate Q-matrix values generated by TESS3 on a geographic map



LaCava et al. 2021

What to do next?

Often good practice to compare results with STRUCTURE/ADMIXTURE and other methods

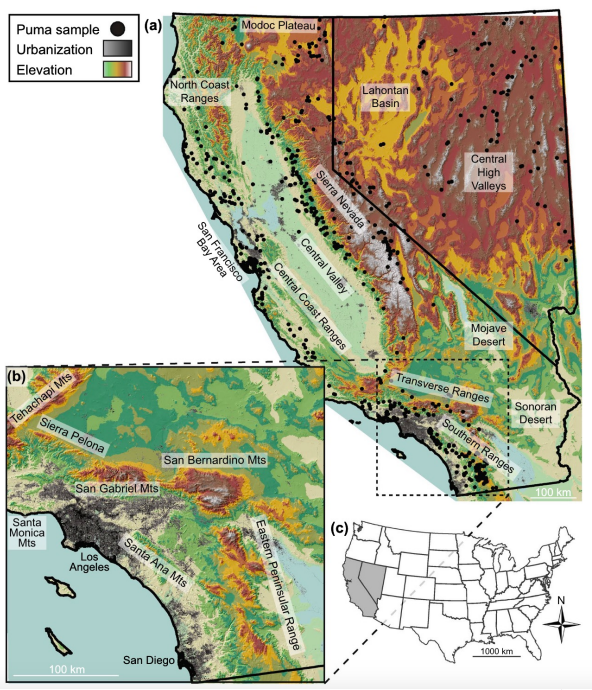
Can also be useful when determine the environmental factors most strongly influencing population structuring

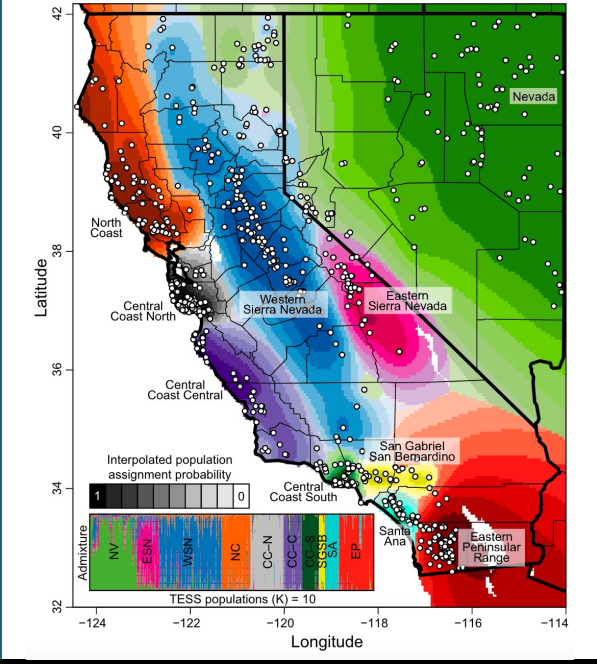
What to do next?

Example: Gustafson et al. 2018

- Genetic data from *Puma concolor* in California and Nevada
  - Evaluated population structure using TESS

- Identified 10 populations
  - Genetic structure was likely influenced by urbanization
- Lower genetic diversity along western coast
  - In some populations, genetic diversity was as low as what is seen for FL panthers





TESS3: Identifying Outlier Loci

Ancestry coefficients can also be useful when assessing adaptive evolution

- TESS3 can use ancestry coefficients to generate Fst values
- Can then generate p-values to identify outliers
  - Benjamini-Hochberg algorithm used to control for false-discovery rate

Takeaways

Assessing population structure is useful for understand how neutral evolutionary processes may have influenced genetic variation

- Can be useful for generating hypotheses and identifying patterns
- Important to account for when assessing adaptive evolutionary processes

TESS3 is a fast method for estimating ancestry coefficients when using thousands of loci and/or individuals

- However, it's usually good to compare results with other methods